DAMEWARE
(Data Mining & Exploration Web Application and Resources)

on behalf of the DAMEWARE collaboration:

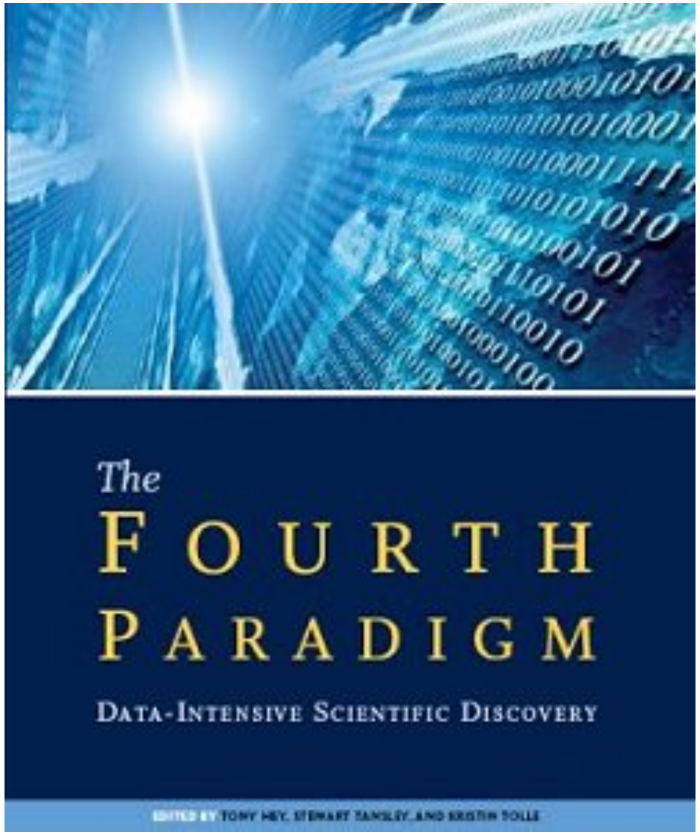
Co PI's:
Massimo Brescia – INAF-OACN (Italy)
George S. Djorgovski – Caltech (USA)
Giuseppe Longo – UNINA (Italy)

Members of the team:

Stefano Cavuoti, Francesco Esposito, Mauro Garofalo,
Marisa Guglielmo, Alfonso Nocella, Francesco Manna – Unina (Italy)
Ciro Donalek, Matthew J. Graham, Ashish A. Mahabal – Caltech (USA)
Raffaele D'Abrusco - CfA Harvard (USA)
Michelangelo Fiore - LAAS (FR)

Pucon, August 2013

Why Machine learning ?



The four legs of modern science

1. **Experiment** (ca. 3000 yrs)

2. **Theory** (few hundreds yrs) mathematical description, theoretical models, analytical laws (e.g. Newton, Maxwell, etc.)

3. **Simulations** (few tens of yrs) Complex phenomena

4. **Data-Intensive science** (now!!!)

http://research.microsoft.com/fourthparadigm/

"***One of the greatest challenges for 21st−century science is how we respond to this new era of data intensive science***"

ASTRONOMY CAN BENEFIT FROM WHAT HAPPENS ELSEWHERE

As a result of large surveys, astronomy has entered an era where

- ***Most data will never be seen by humans!***

  The need for data storage, network, database-related technologies standards, etc.

- ***Most knowledge hidden behind data complexity is potentially lost***

  Most (if not all) empirical relationships known so far depend on 3 parameters …. (e.g. fundamental plane of E galaxies and bulges). Simple universe or rather human bias?

- ***Most data (and data constructs) cannot be comprehended by humans directly!***

  The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/Machine-assisted discovery

The various components of the data challenge

Data Gathering (e.g., from sensor networks, telescopes…)

Data Farming:

- Storage/Archiving
- Indexing, Searchability
- Data Fusion, Interoperability

Specific projects (pipelines, etc.) & Virtual observatory

Data Mining (or Knowledge Discovery in Databases):

- Pattern or correlation search
- Clustering analysis, automated classification
- Outlier / anomaly searches
- Hyperdimensional visualization.

Data understanding

- Computer aided understanding
- KDD
- Etc.

New Knowledge

Uncharted land

DAMEWARE

Is a web−based application (FREE AND OPEN TO THE PUBLIC) for massive data mining based on a suite of machine learning methods on top of a virtualized hybrid computing infrastructure

A joint effort between University Federico II, INAF−OACN & Caltech, aimed at implementing (as web 2.0 apps and services) a scientific gateway for data exploration on top of a virtualized distributed computing environment

Alfa released NOW

http://dame.dsf.unina.it/

- Science and management
- Technical documents
- Template science cases
- Newsletters
- Tutorials

Effective DM requires complex work-flows

Use case
→ pre-processing
→ feature selection
→ choice of DM model
→ experiments
→ evaluation of results

The logic behind DAMEWARE

Use case

Functionality

- Classification
- Regression
- Clustering
- Feature selection

DM models

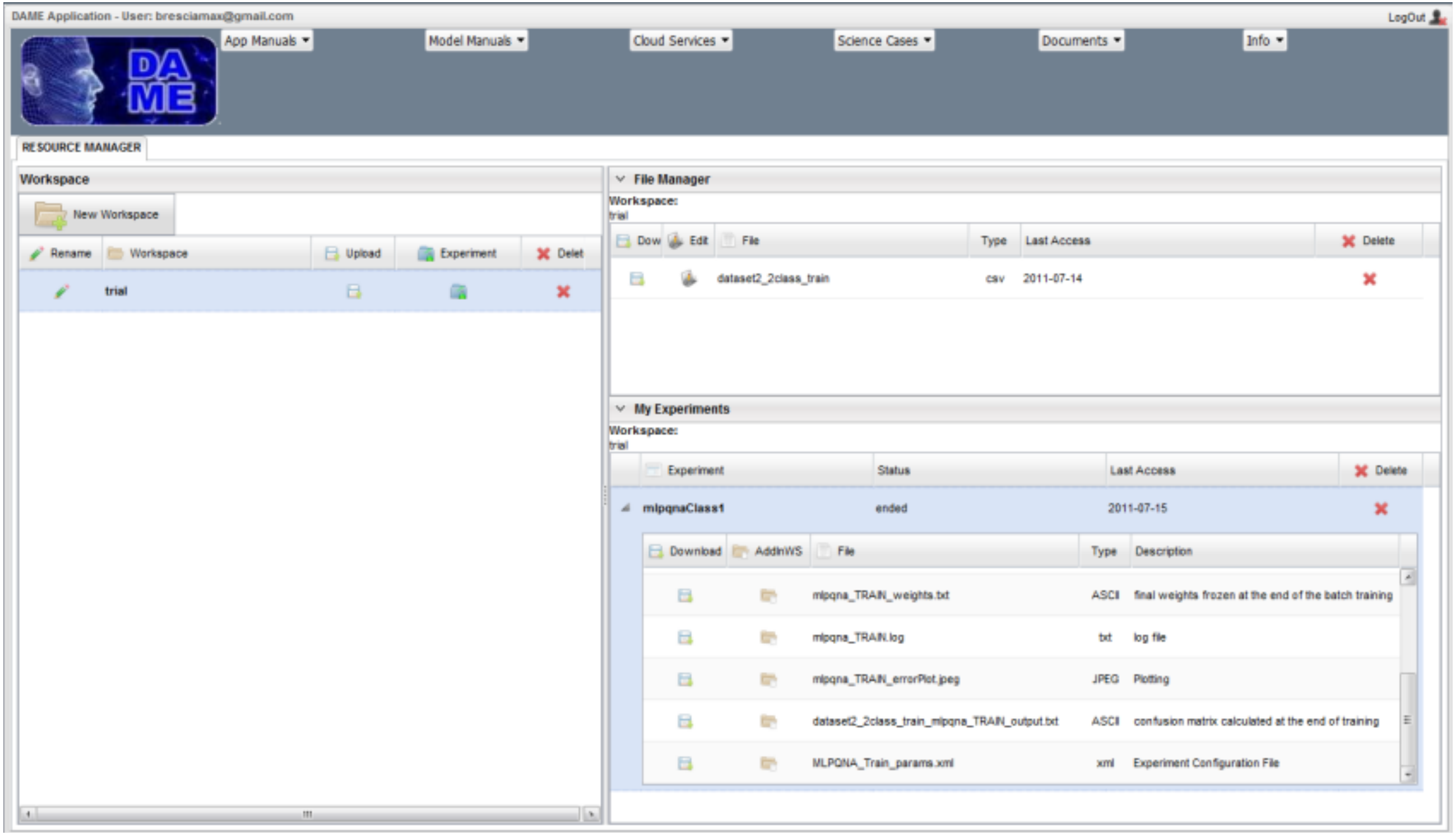
GAME	S, C,R
MLPBP	S, C,R
MLPGA	S, C,R
MLPQNA	S, C,R
SVM	S, C,R
K-Means	U, Cl
ESOM	U, Cl
SOFM	U, Cl
SOM	U, Cl
PPS	U, Cl, FS

Experiments

1-st
2-nd
3-rd
4-th
….

N-th

DAMEWARE the GUI

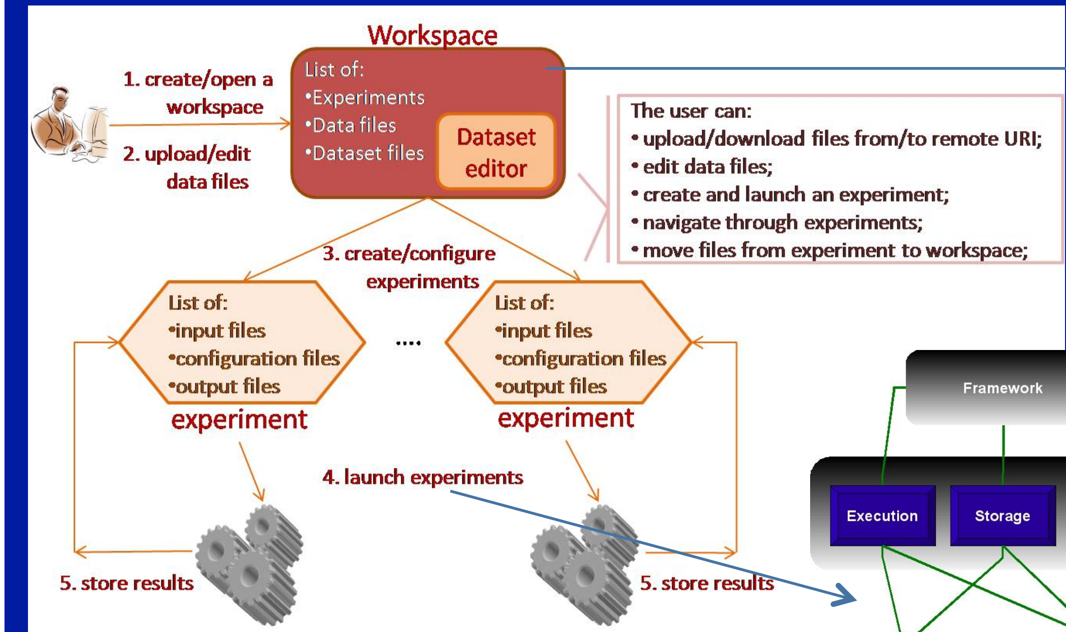


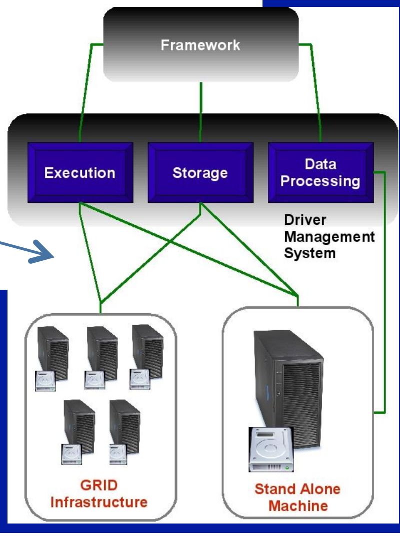
DAMEWARE

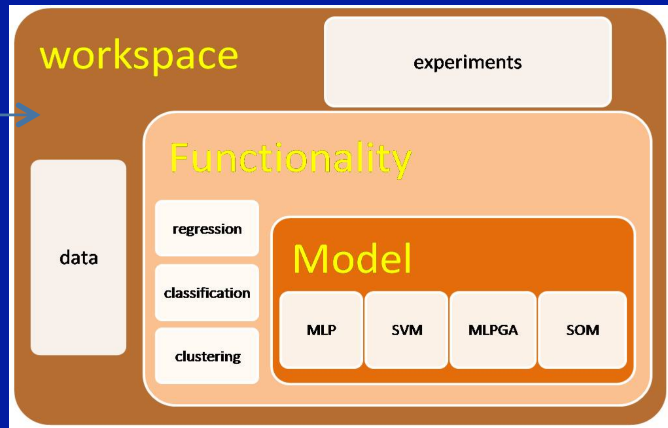
It is multi-disciplinary platform (astronomy, bioinformatics and medical diagnostics)

End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone…)

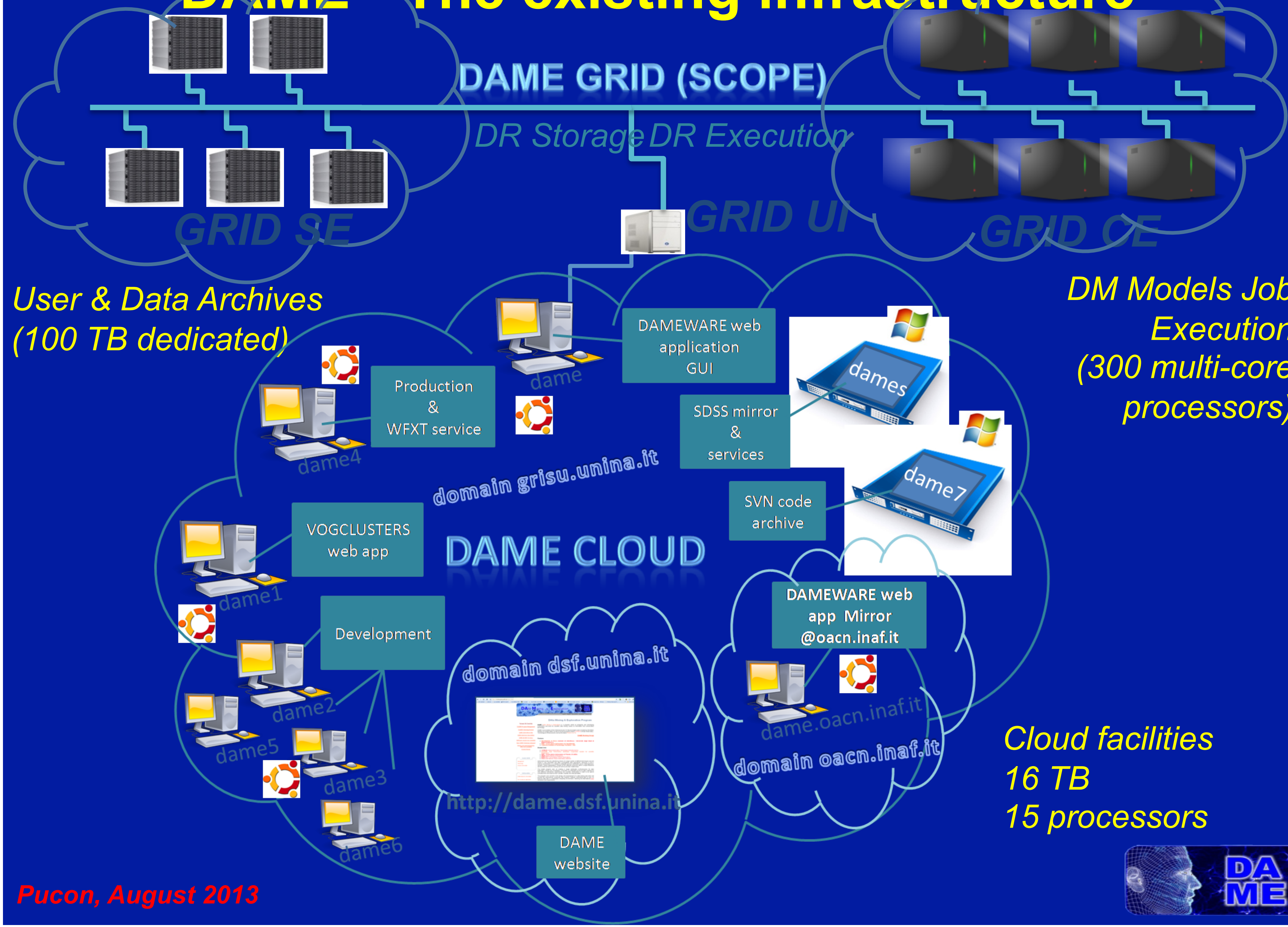
User can automatically plug-in his/her own algorithm and launch experiments through the Suite via a simple web browser







DAME - The existing infrastructure



User & Data Archives (100 TB dedicated)

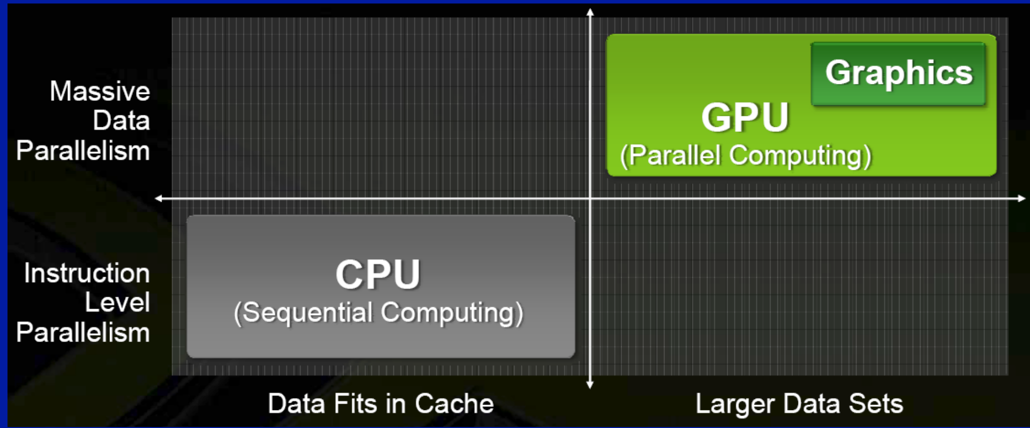
DM Models Job Execution (300 multi-core processors)

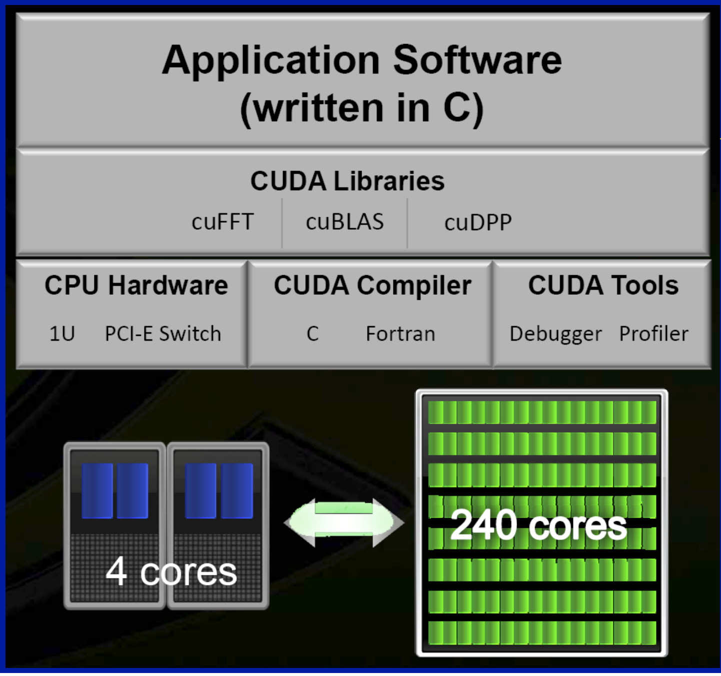
Cloud facilities
16 TB
15 processors

GPU technology ... sometime useful

The Graphical Processing Unit is specialized for compute-intensive, highly parallel computation (exactly what graphics rendering is about).

« GPU have evolved to the point where many real world apps are easily implemented on them and run significantly faster than on multi-core systems.»





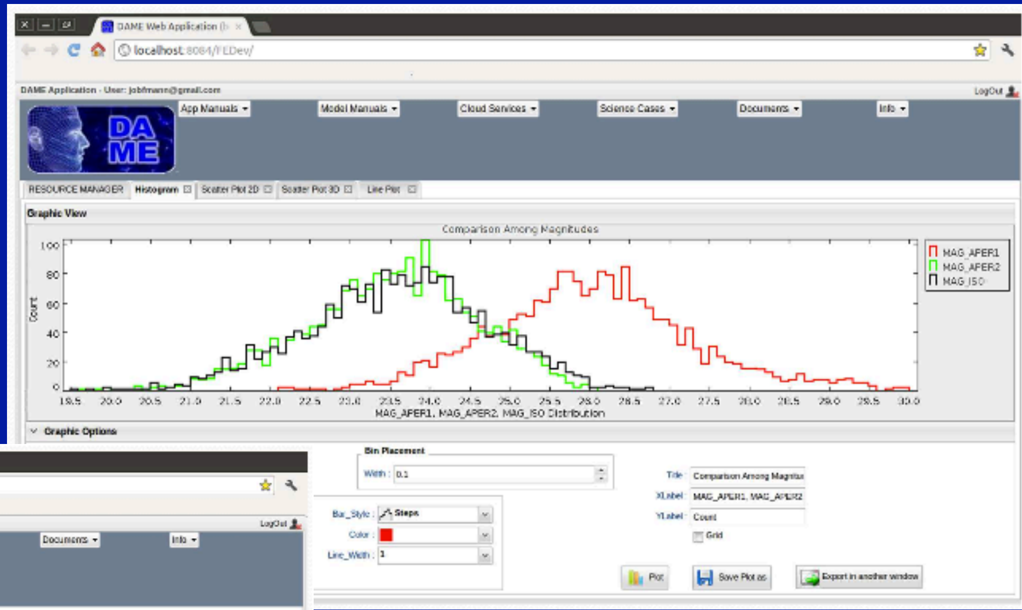
DAME - GAME
Genetic Algorithm Mining Experiment

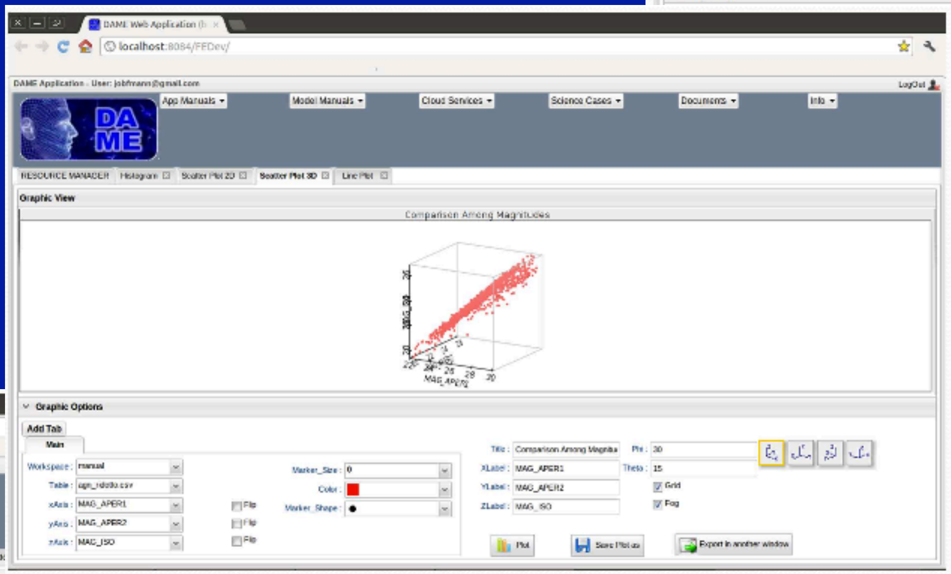
GAME is a pure genetic algorithm developed in order to solve supervised problems of regression or classification, able to work on Massive Data Sets (MDS).

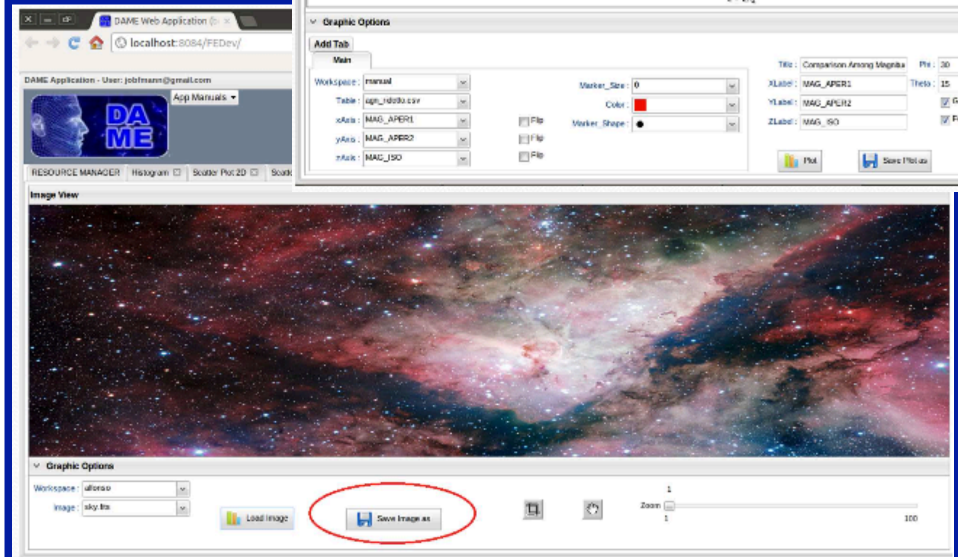
Graphical capabilities in DAMEWARE

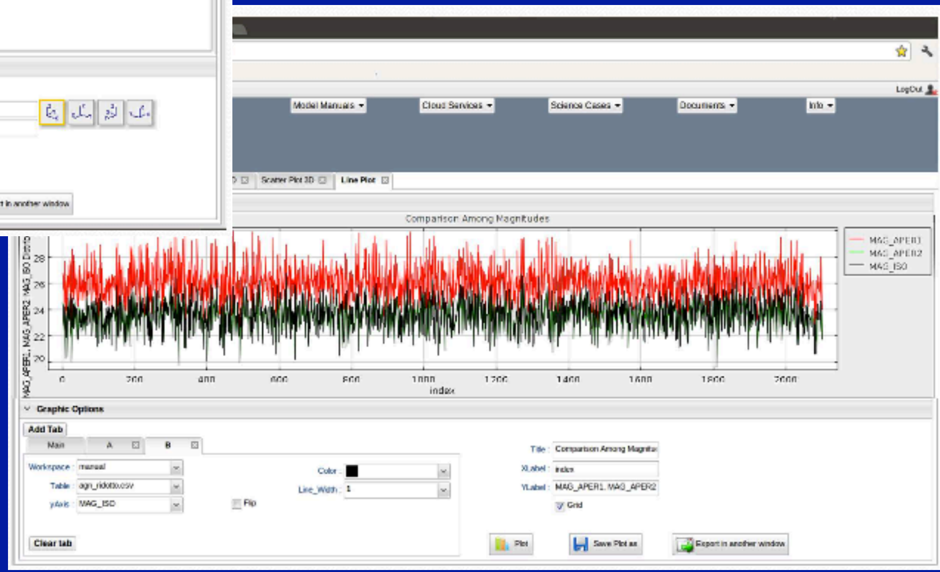
- Histograms
- 2-D & 3-D plots
- Line plots
- Image visualization

Java client









AGN identification and classification

Photometric classification of emission line galaxies with Machine Learning methods, Cavuoti et al., 2013, MNRAS, submitted

Star/Galaxy separation

The detection of globular clusters as a data mining problem, Brescia et al., 2012, MNRAS, 421, 1155-1165 (arXiv:1110.2144)

GPUs for astrophysical data mining. A test on the search for candidate globular clusters in external galaxies. S. Cavuoti, et al., New Astronomy, april 20, 2013, http://dx.doi.org/10.1016/j.newast.2013.04.004 (astro-ph: 1304.0597)

Photometric redshifts

Mining the SDSS archive. I. Photometric redshifts in the nearby universe, D'Abrusco, Logno G., Walton N., 2007, ApJ, 663, 752

Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation , O. Laurino, R. D'Abrusco, G. Longo, and G. Riccio, MNRAS, 2011, 418, 2165 (arXiv/1107.3160);

Photometric redshifts with Quasi Newton Algorithm (MLPQNA) Results in the PHAT1 context, Cavuoti et al. 2012, , Astronomy and Astrophysics 546, 13, (ArXiv:1206.0876)

Photometric redshifts for quasars in multiband surveys, M. Brescia et al., 2013, ApJ, 772, 140 (astro-ph: 1305.5641)

Inside catalogs: a comparison of source extraction software, M. Annunziatella, et al., 2012, PASP, 125, 68 (astro-ph:1212.0564).

Other

Astroinformatics, data mining and the future of astronomical research, M. Brescia & G. Longo, 2012, invited to appear in proceed. of IFDT2 - 2nd International conference frontiers on diagnostic technologies (arXiv:1201.1867)

CLASPS: a new methodology for knowledge extraction from complex astronomical data sets, R. D'Abrusco, G. Fabbiano, S.G. Djorgovski, C. Donalek, O. Laurino & G. Longo, 2012, ApJ, 755, 92 (ArXiv: 1206.2919)

Current Applications in other fields

Medical diagnosis of alzhaimer
(S. Cocozza et al.)

Brain tomography analysis
(Bellotti M. et al.)

Real time classification of ethernet data flows
(G. Ventre et al.)

Etc…

Some statistics (2013)

Ca. 100 users, > 12.000 experiments

An operative example

Use case:
Photometric redshifts evaluation for quasars in panchromatic surveys

Functionality: regression

Pre-processing: preparation of KB (10^5 objects)
removal of NaN,
splitting of train, validation, test sets

Feature selection (>50 experiments)

Selection of best DM model: SVM; MLPBP, MLP-GA, GAME, MLPQNA

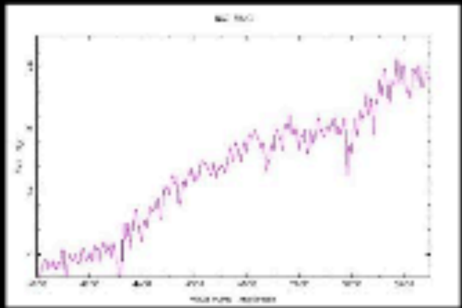
Training
Validation (10 experiments)
Test

TOTAL of ca. 2000 experiments

Visualization, comparison & Evaluation of results

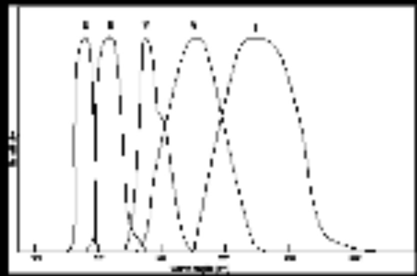
PHOTOMETRIC REDSHIFTS AS A INVERSE PROBLEM

Spectral Energy Distribution convolved with band filters



Galaxy spectrum - $F(\lambda)$

X



Photometric system - $S_i(\lambda)$

=

$$m_U = -2.5\log_{10}\frac{\int F(\lambda)S_U(\lambda)d\lambda}{\int S_U(\lambda)d\lambda} + c_u$$

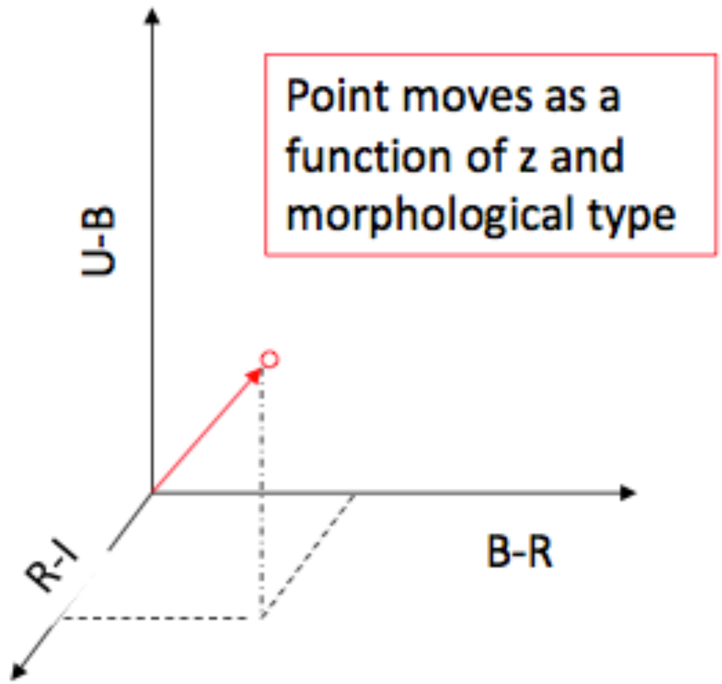
$$m_B = -2.5\log_{10}\frac{\int F(\lambda)S_B(\lambda)d\lambda}{\int S_B(\lambda)d\lambda} + c_B$$

Color indexes

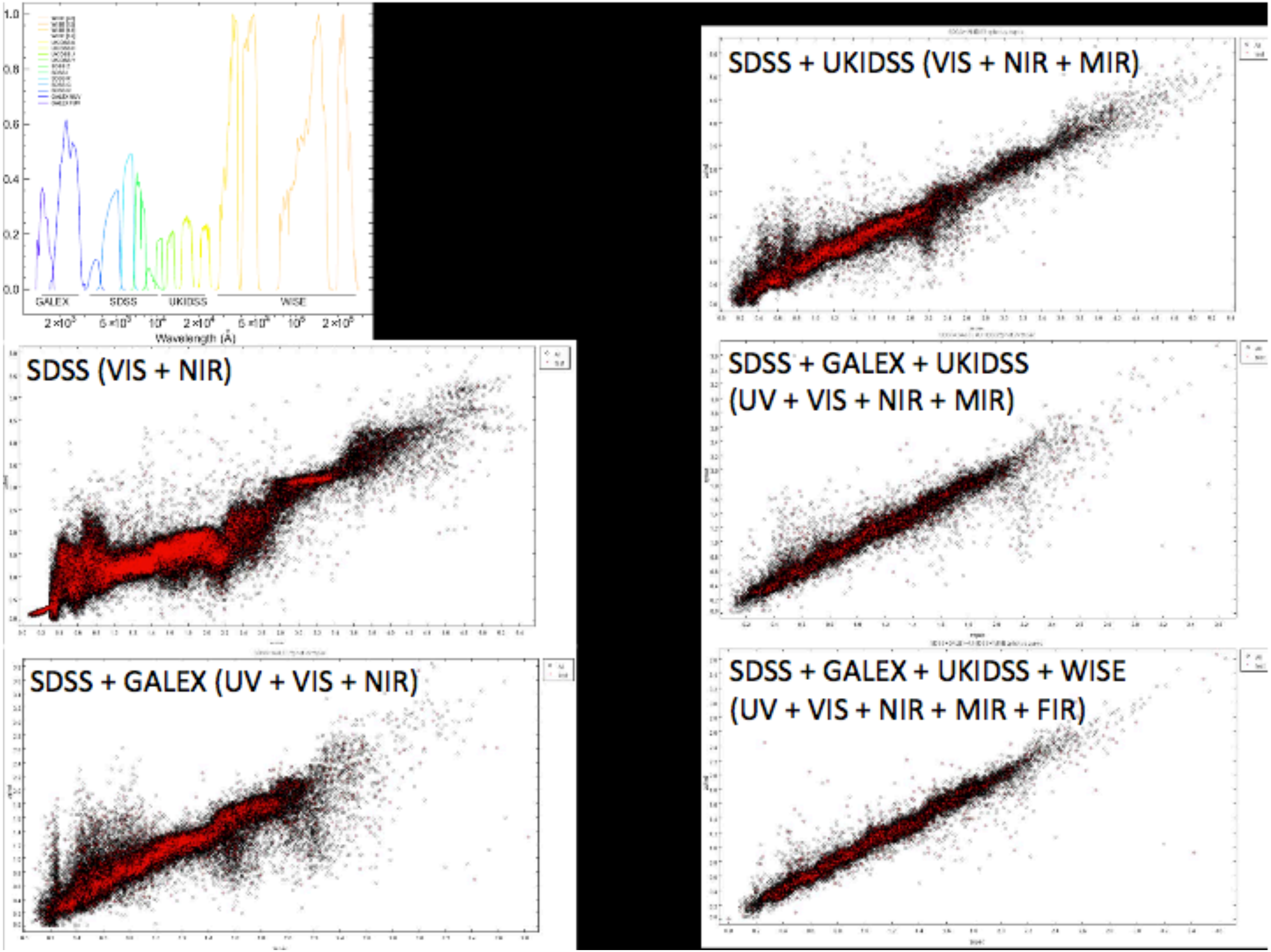
$U - B \equiv m_U - m_B$

$B - R \equiv m_B - m_R$

etc.



Phot-z are an inverse problem



GALEX
SDSS
UKIDSS
WISE
Wavelength (Å)
SDSS (VIS + NIR)
SDSS + GALEX (UV + VIS + NIR)
SDSS + UKIDSS (VIS + NIR + MIR)
SDSS + GALEX + UKIDSS
(UV + VIS + NIR + MIR)
SDSS + GALEX + UKIDSS + WISE
(UV + VIS + NIR + MIR + FIR)

TEST	MEAN	σ	out. 1σ	out. 2σ	out. 3σ	out. 4σ	TOTAL OBJECTS
E5	0,0005	0,118	18,67%	4,01%	1,51%	0,87%	3787
E16	-0,0004	0,154	18,11%	4,75%	1,98%	0,98%	3787

Table 3. Summary of the statistical indicators alread used in Table xx (bias, σ and the percentage of outliers at, respectively, 1,2,3 and 4 σ computed as in citebovy2012 on all objects (test and training set).

ZSPEC BIN	EXP	BIAS	SIGMA	$\lvert\Delta Z\rvert > 0.1$	$\lvert\Delta Z\rvert > 0.2$	$\lvert\Delta Z\rvert > 0.3$	$\lvert\Delta Z\rvert > 0.4$	OBJECTS
TRAIN Only								
[0.2, 1.0]	E23	-0.0897	0.206	44.94%	22.78%	13.29%	8.86%	316
[0.2, 1.0]	E5	-0.0183	0.118	27.53%	7.28%	2.22%	1.27%	316
[0.2, 1.0]	E16	-0.029	0.127	29.43%	10.76%	3.48%	1.90%	316
[0.2, 1.0]	E10	-0,1807	0,281	64,87%	39,87%	26,58%	18,67%	316
[1.4, 3.0]	E23	0.1209	0.273	59.05%	32.33%	21.98%	14.22%	232
[1.4, 3.0]	E5	0.0364	0.18	38.36%	14.66%	8.19%	4.74%	232
[1.4, 3.0]	E16	0.0408	0.183	40.09%	15.52%	8.62%	4.74%	232
[1.4, 3.0]	E10	0,2188	0,367	62,50%	41,38%	28,88%	22,84%	232
TRAIN+TEST								
[0.2, 1.0]	E23	-0.0911	0.23	46.24%	23.18%	13.77%	9.03%	1583
[0.2, 1.0]	E5	-0.0174	0.101	21.04%	4.17%	1.58%	0.82%	1583
[0.2, 1.0]	E16	-0.0326	0.142	30.01%	9.85%	4.67%	2.65%	1583
[0.2, 1.0]	E10	-0,1877	0,287	63,93%	39,55%	27,48%	19,08%	1583
[1.4, 3.0]	E23	0.1238	0.269	56.24%	30.74%	18.69%	12.49%	1145
[1.4, 3.0]	E5	0.0271	0.139	31.44%	9.61%	3.93%	2.10%	1145
[1.4, 3.0]	E16	0.0492	0.183	39.83%	14.93%	7.95%	4.37%	1145
[1.4, 3.0]	E10	0.2488	0,37	64,28%	44,02%	32,23%	24,72%	1145

The new mantra

Discovery of rare and unknown…

Search for higher order correlations etc…

The new mantra

Discovery of rare and unknown…

Search for higher order correlations etc…

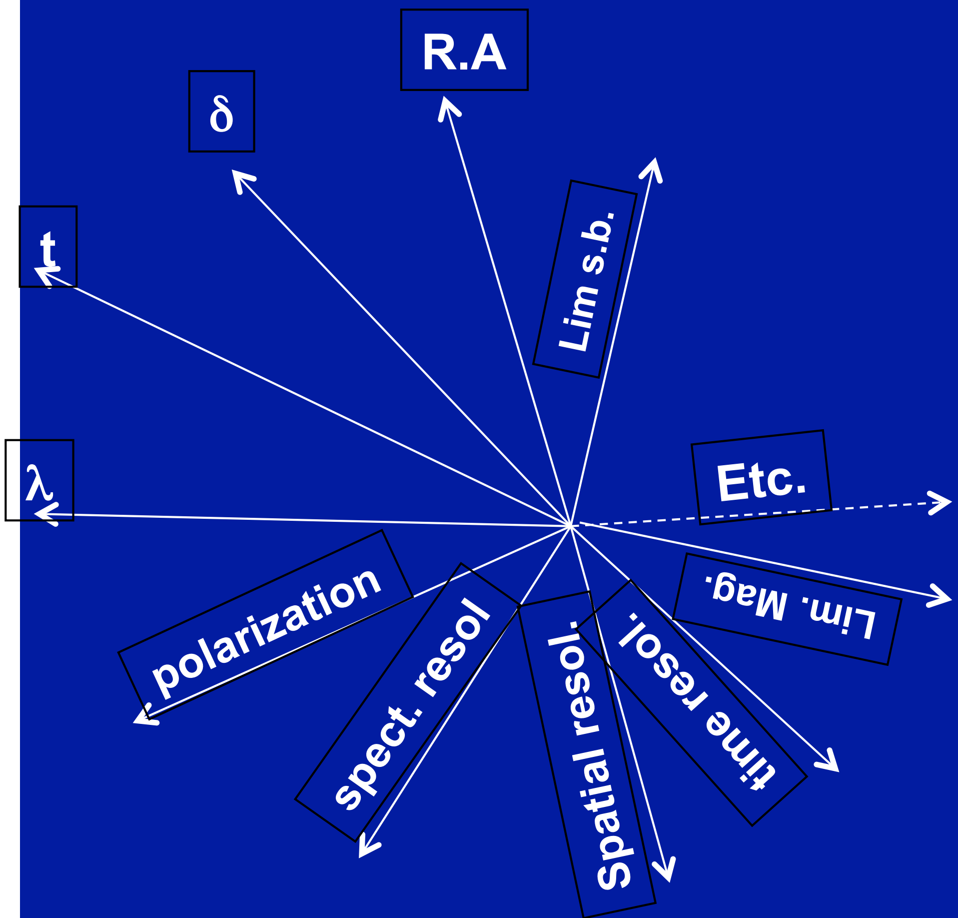
Small data sets: serendipity or guided luck

Large Data sets: clustering….

The new mantra

Discovery of rare and unknown…

Search for higher order correlations etc…



Each datum is defined by n measured parameters.

- X,y,t
- Flux
- Polarization
- wavelength
- Etc..

New sensor

$$p \in \Re^N \qquad N >> 100$$

Exploration of PS with $N > 10^9$, $D \gg 100$, $K > 10$ Is anything but simple

N = no. of data vectors,
D = no. of data dimensions
K = no. of clusters chosen,
K_{max} = max no. of clusters tried
I = no. of iterations, M = no. of Monte Carlo trials/partitions

K-means: $K \times N \times I \times \mathbf{D}$
Expectation Maximisation: $K \times N \times I \times \mathbf{D^2}$
Monte Carlo Cross-Validation: $M \times K_{max}^2 \times N \times I \times \mathbf{D^2}$
Correlations ~ $N \log N$ or N^2, ~ D^k ($k \geq 1$)
Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)
SVM > ~ $(NxD)^3$

→ **Lots (…truly lots and lots…) of computing power**

Moving programs not data: the true bottle neck

Data Mining + Data Warehouse =
Mining of Warehouse Data

- For organizational learning to take place, data from must be gathered together and organized in a consistent and useful way – hence, Data Warehousing (DW);
- DW allows an organization to remember what it has noticed about its data;
- Data Mining apps should be interoperable with data organized and shared between DW.

Interoperability scenarios

DA1 **Data+apps Exchange** DA2

Full interoperability between DA (Desktop Applications)
Local user desktop fully involved (requires computing power)

NO MDS

DA **Data+apps Exchange** WA

Full WA → DA interoperability
Partial DA → WA interoperability (such as remote file storing)
MDS must be moved between local and remote apps
user desktop partially involved (requires minor computing and storage power)

NO MDS

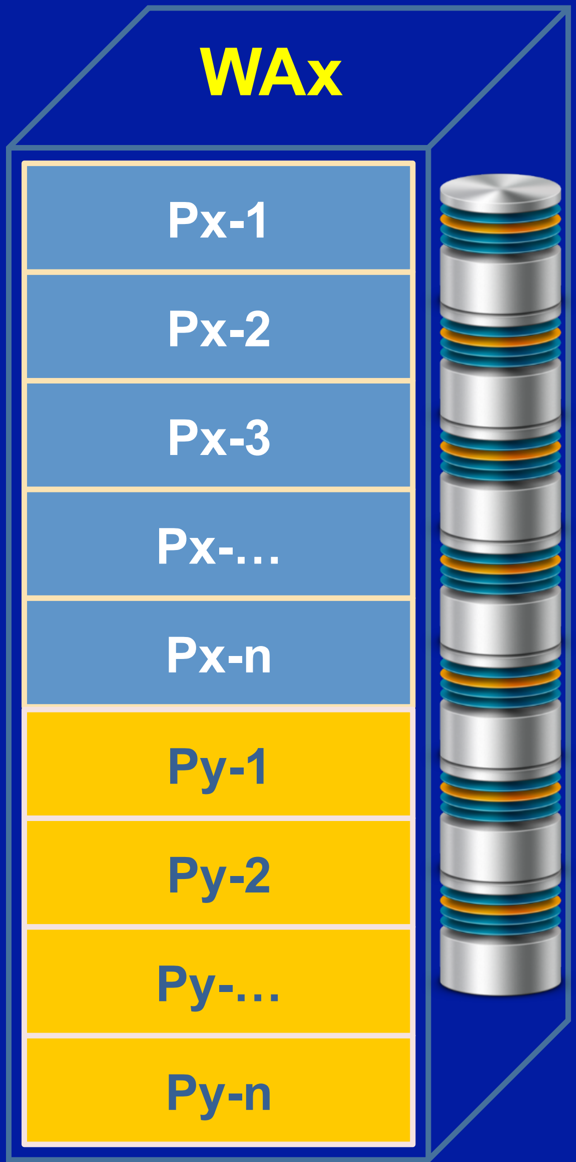
WA **Data+apps Exchange** WA

Except from URI exchange, no interoperability and different accounting policy
MDS must be moved between remote apps (but larger bandwidth)
No local computing power required

NO MDS

The Lernaean Hydra KDD

After a certain number of such iterations…



The scenario will become:

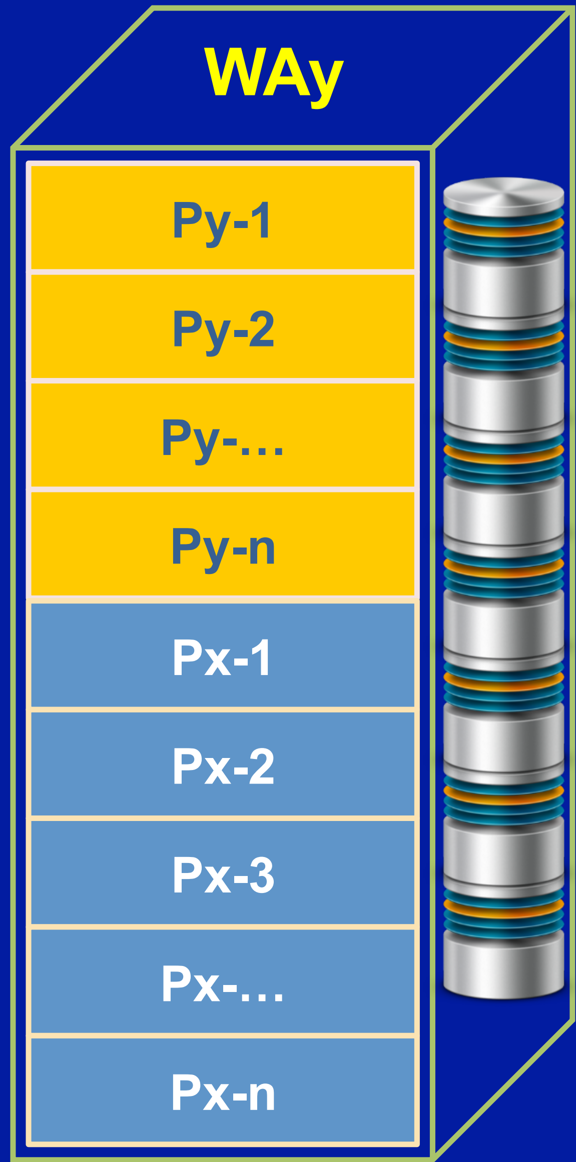
No different WSs, but simply one WS with several sites (eventually with different GUIs and computing environments)

All WS sites can become a mirror site of all the others

The synchronization of plugin releases between WSs is performed at request time

Minimization of data exchange flow (just few plugins in case of synchronization between mirrors)

YES MDS!

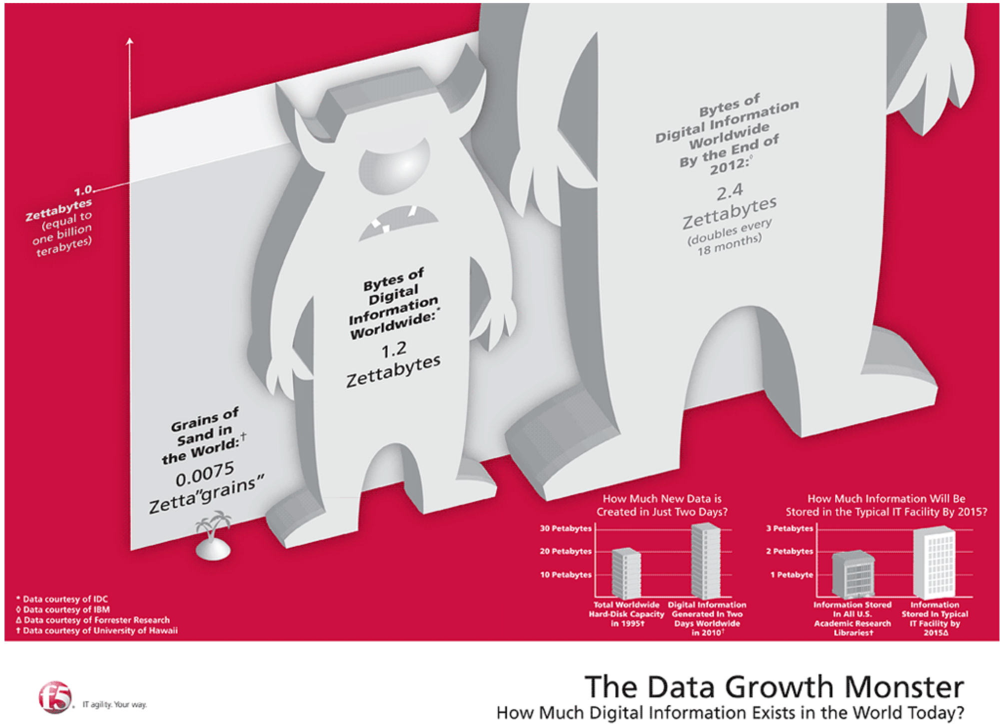


The Data Growth Monster
How Much Digital Information Exists in the World Today?

astronomical problems are a piece of cake….

Growth of digital data worldwide (2012)

1 ZB/yr or = 10^9 Terabyte



astronomical problems are a piece of cake….

Growth of digital data worldwide (2012)

1 ZB/yr or = 10^9 Terabyte

NSA listening station in Bluffdale, South Dakota

1 YB of storage = 10^{12} TB

Indexed, searched, mined…

With mainly unknown technology which will slowly leak out to the scientific community


Thanks for the attention …
… and to Eduardo for organizing the meeting …